

Understanding Emotional Responses to Mobile Video Advertisements via Physiological Signal Sensing and Facial Expression Analysis

Phuong Pham^{1,2}, Jingtao Wang^{1,2}

¹Department of Computer Science, ²Learning Research and Development Center (LRDC)
University of Pittsburgh, PA, USA
{phuongpham, jingtaw}@cs.pitt.edu

ABSTRACT

Understanding a target audience's emotional responses to video advertisements is crucial to stakeholders. However, traditional methods for collecting such information are slow, expensive, and coarse-grained. We propose AttentiveVideo, an intelligent mobile interface with corresponding inference algorithms to monitor and quantify the effects of mobile video advertising. AttentiveVideo employs a combination of implicit photoplethysmography (PPG) sensing and facial expression analysis (FEA) to predict viewers' attention, engagement, and sentiment when watching video advertisements on unmodified smartphones. In a 24-participant study, we found that AttentiveVideo achieved good accuracies on a wide range of emotional measures (the best average accuracy = 73.59%, kappa = 0.46 across 9 metrics). We also found that the PPG sensing channel and the FEA technique are complimentary. While FEA works better for strong emotions (e.g., joy and anger), the PPG channel is more informative for subtle responses or emotions. These findings show the potential for both low-cost collection and deep understanding of emotional responses to mobile video advertisements.

Author Keywords

Heart Rate; Computational Advertisement; Physiological Signal; Affective Computing; Mobile Interfaces.

ACM Classification Keywords

H.5.2 [Information interfaces and presentation]: User Interfaces – graphical user interfaces, theory and methods.

INTRODUCTION

In 2015, U.S. online advertising revenues reached \$59.6 billion, wherein digital video generated \$4.2 billion [39]. Mobile advertising is the fastest growing segment of this

domain (e.g. over 66% growth vs. a 23% growth for the entire industry in 2014 [39]). Despite the huge revenues and the rapid growth, it is still challenging to evaluate the *quality* of advertising. For example, the efficacy of *direct response advertising* [5], i.e. persuading a prospective customer to purchase specific merchandise, can be quantified through metrics such as click-through-rate (CTR) [25, 46], conversion ratio (CVR) [20, 49], and cost per click (CPC) [11, 49]. It is much more challenging to measure the effectiveness of *branding advertising* [26]. Since branding advertising intends to increase customers' awareness, trust, and sometimes loyalty towards a brand, there are limited short-term user behaviors that can be observed and analyzed.



Figure 1. AttentiveVideo with dual video controls (top: touch widgets for non-ad video watching; bottom: on-lens finger gestures from the back camera and facial tracking from the front camera for advertisement watching).

Self-report data, focus groups, and behavior analysis are commonly used to evaluate *branding advertisements* [1, 22, 38]. However, these methods are expensive, time-consuming, and may not lead to reliable results due to the inherent ambiguity in reporting viewers' subjective feelings. Meanwhile, autonomic feedback techniques, such as facial expression analysis [15, 28] and physiological signals [19, 31], could serve as orthogonal dimensions to understand prospective customers' emotional responses to advertisements (ads) [31]. Unfortunately, most autonomic

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

IUI 2017, March 13–16, 2017, Limassol, Cyprus

© 2017 ACM. ISBN 978-1-4503-4348-0/17/03...\$15.00

DOI: <http://dx.doi.org/10.1145/3025171.3025186>

feedback techniques require either dedicated sensors [19, 31] or PCs connected to high-speed internet [15, 28] to run.

To address these challenges, we propose AttentiveVideo, a lightweight mobile video player that can collect users' emotional responses to mobile ads via two modalities on unmodified smartphones. AttentiveVideo utilizes a dual video control system (Figure 1). For regular video materials (e.g. movies and TV shows), AttentiveVideo is similar to today's mobile video apps and uses on-screen touch widgets for play and pause. When a subsidized advertisement video is playing, AttentiveVideo enables and requires video control through *on-lens* finger gestures, i.e. covering and holding the back camera lens to play the ad, while uncovering the lens to pause the ad. As a byproduct of this tangible video control paradigm, the viewer's photoplethysmography (PPG) signals are extracted implicitly by continuously monitoring her fingertip's transparency changes through the back camera. Furthermore, AttentiveVideo captures and analyzes the viewer's facial expressions through the front camera in parallel. In this paper, we show that it is feasible to predict viewers' *attention*, *engagement*, and *sentiment* responses to mobile video advertisements with good accuracies on unmodified smartphones via a combination of implicit PPG sensing and facial expression analysis. Major contributions of this paper include:

- The design, implementation, and evaluation of an intelligent mobile interface, AttentiveVideo, for the automatic collection of emotional responses to mobile video ads on unmodified smartphones.
- A direct comparison of two rich modalities that sense human affect, i.e. PPG signals and facial expressions, in the context of mobile advertising.
- We show that a wide range of viewer's emotional responses to video advertisements can be inferred with higher accuracy by combining the complementary results from PPG sensing and facial expression analysis.

RELATED WORK

Advertising Effectiveness

During the past 20 years, advertisements, a.k.a. commercials, have shifted from uniform presentations [25] to more personalized and relevant advertisements [7, 24, 30] via techniques such as *behavior targeting*. Researchers have improved the effectiveness of *direct response advertising* [5] by identifying crucial factors and empirical techniques, such as item-based collaborative filtering [24], named entities recognition [7], relevant embedded positions [30], and animation in ad banners [25]. Large scale, data driven approaches optimizing short-term behavior metrics, such as CTR [25, 46], CVR [20, 49] and viewing duration [22], are becoming the de-facto standard for evaluating the efficacy of direct response advertising.

Meanwhile, the evaluation and improvement of *branding advertising* is still an open problem. Self-report and polling [1, 31, 38] are the most popular techniques to date. However, these technologies require additional cognitive workload in reporting one's emotional responses to ads. Meanwhile, researchers have explored the use of autonomic feedback channels, e.g. skin conductance [31], heart rate [19, 31], and facial expressions [28], to measure participants' subjective feelings towards an advertisement. However, most of the existing research efforts have focused on the feasibility in lab settings and PC environments. The requirements on dedicated wearable sensors and a high-speed network prevent the wide adoption of these approaches in everyday settings, especially in mobile environments. In comparison, we show that it is possible to leverage the front camera and the back camera on smartphones to capture and infer a rich set of viewer's emotional responses to video advertisements without any hardware modifications.

Affective Computing

Affective computing [37] refers to the design and evaluation of computerized techniques that can recognize, interpret, and respond to human affective states. Affect detection has been studied in various contexts such as education [6], human-robot communication [40], and healthcare [18].

Physiological signals, such as skin conductance [6], heart rate variability (HRV) [18], and acoustic-prosodic features [40] have been leveraged as informative input modalities to detect users' affective states. Lang [19] explored the feasibility of using heart rate to detect physiological arousal evoked by TV ads. Please refer to Calvo and D'Mello [8] for a comprehensive survey on the use of physiological signals in affect/emotion detection.

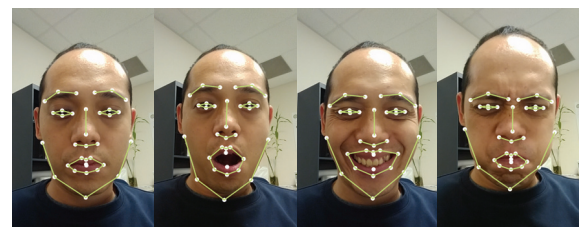


Figure 2. Example faces showing 26 landmarks detected.

In addition to physiological signals, McDuff et al. [28, 29] recently explored the use of facial expression analysis (FEA) to analyze viewers' emotional responses and purchase intent from video advertisements on PCs via Adobe Flash widgets embedded in web pages. Similarly, FEA has been explored by researchers in the affective computing community and the computer vision community to understand human emotions [8]. Most of today's FEA systems first detect landmark points in a face via facial alignment (Figure 2), and then map the temporal changes of

landmark points to emotions via the Facial Action Coding System (FACS)¹.

Previous work [3, 10, 17] has shown improvements in accuracy when combining facial expressions and physiological signals into a multimodal system. Bailenson et al. [3] found the addition of physiological features to facial expressions increased the prediction precision of sadness by over 15% and of amusement by 9%. Hussian et al. [17] achieved higher accuracies in valence and arousal prediction when using ensemble methods for facial-based and physiological-based modalities. D’Mello and Graesser [10] showed an improvement in their facial-physiological system when detecting learners’ cognitive and affective states.

One common requirement of these multi-modal detection approaches is the additional sensors for data collection. The extra costs in purchasing sensors and additional efforts in carrying them around limit the scalability of such systems. AttentiveLearner [35, 42] is perhaps the most relevant to our research. AttentiveLearner [35, 42] can capture users’ physiological signals on today’s smartphones without any hardware modification. Both AttentiveLearner and our research use on-lens finger gestures for tangible video control and PPG sensing on unmodified smartphones. However, there are two major differences between the two systems. First, AttentiveLearner focuses on watching lecture videos in Massive Open Online Courses (MOOCs) and flipped classrooms, whereas AttentiveVideo is optimized for monitoring mobile video advertisements. Compared with lecture videos, advertisements are much shorter (5 – 20 minutes vs. 30 seconds) and usually carry stronger stimuli to elicit emotional responses from the audience. As detailed in follow-up sections, such differences impose significant challenges to algorithm design; Second, AttentiveLearner focuses on the collection and use of PPG sensing from the back camera alone. In contrast, our research explores the comparison and combination of both the PPG sensing channel via the back camera and the facial expressions from the front camera on unmodified devices. To the best of our knowledge, AttentiveVideo is the *first* mobile system to predict users’ affective states via a combination of PPG sensing and FEA in real time on today’s unmodified smartphones.

Mobile Video Interfaces

Being the fastest growing segment among online advertising, mobile advertising is a promising paradigm for advertisers [39]. Researchers have proposed various interaction techniques to streamline interaction with videos on mobile devices. Ganhör [12] divided a phone’s screen into four panes and used one or more panes for efficient video navigation. Wu et al. [41] studied one-handed tilting and shaking gestures for video browsing. Zhang et al. [48]

utilized the touch screen for both video navigation and collaborative sketches. These approaches focused more on video control on mobile devices rather than collecting and understanding user emotional responses. By comparison, AttentiveVideo uses two modalities, i.e. PPG signals from the back camera and facial expressions from the front camera, to enable an emotion-aware interface to infer emotional responses to mobile video advertisements.

DESIGN OF ATTENTIVEVIDEO

AttentiveVideo is designed as a video player for mobile devices. We expect end-users will use AttentiveVideo to watch *copyrighted* yet *ad-subsidized* movies or TV shows on their smartphones. AttentiveVideo has three unique features when compared to existing mobile video players: 1) the dual video control interface, i.e. *on-screen* UI widgets for controlling regular videos and *on-lens* finger gestures for controlling ads, 2) an autonomic feedback collection interface, and 3) affect inference algorithms.

Dual Video Control Interface

In AttentiveVideo, the *on-screen* UI widgets for controlling the playback of regular videos are similar to existing mobile video players (Figure 1, top).

When it is time to show a sponsored advertisement video, AttentiveVideo requires a viewer to cover-and-hold the lens of the back camera with her finger for the duration of the advertisement (Figure 1, bottom) while uncovering the camera lens pauses the ad. AttentiveVideo extends the *Static LensGesture* algorithm [45] to detect the lens-covering actions. This algorithm achieved an accuracy of 99.59% for video control [42]. At the same time, the front camera captures and analyzes the viewer’s facial expressions in real-time.

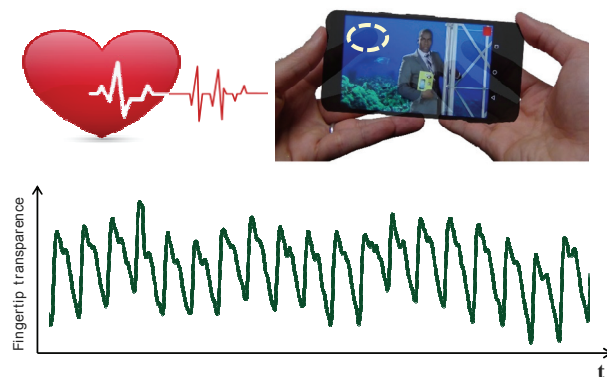


Figure 3. AttentiveVideo captures transparency changes of fingertips implicitly during ad watching via the back camera (top). The PPG waveforms extracted (bottom).

This seemingly “*awkward*” video control mechanism has at least three advantages in the context of subsidized mobile advertising: 1) this mechanism *intentionally* makes it harder for a viewer to skip the sponsored advertisement. Only live body parts (e.g. finger or earlobe) supporting PPG sensing

¹ FACS is an encoding standard for facial expressions, including a total of 47 facial movements.

can be used for lens covering to enable ad playback. Paradoxically, making the ad hard to skip is beneficial to both advertisers and viewers. Advertisers can get increased reception and richer feedback. Consequently, viewers can enjoy more and higher quality video resources supported by advertisers. Meanwhile, viewers always have the freedom to switch to a "pay-per-view" option if they are not interested in ad watching; 2) it provides natural tactile feedback from the bezel of the back camera when a user is holding the phone in landscape mode; and 3) the cover-and-hold gesture allows AttentiveVideo to implicitly capture the user's physiological signals and facial expressions during ad watching. As detailed in follow-up sections, such PPG signals and facial expressions can be used to infer users' emotions during the advertisement and are valuable to advertisers. Xiao and Wang [42] have also found this on-lens finger gesture intuitive and comfortable to use in a series of disciplined usability studies. In this paper, we explore the feasibility of further reducing user effort by minimizing the covering time in a video while keeping the system's accuracy high.

Autonomic Feedback Collection Interface

AttentiveVideo collects a user's PPG signals implicitly by analyzing fingertip transparency changes in real time through the back camera (Figure 2). The underlining mechanism is tied to user's cardiac cycles. In every cardiac cycle, the heart pumps blood to the capillary vessels and fingertips of users. The arrival and withdrawal of fresh blood changes the transparency of fingertips and can be detected by the built-in camera when the user's fingertip is covering the lens of the camera [14]. At the same time, AttentiveVideo utilizes the front camera to capture user's facial expressions while watching an advertisement. As a result, AttentiveVideo enables the automatic collection of both PPG signals and facial expressions *implicitly* during mobile ad watching.

Affect Inference Algorithms

According to previous research [19, 36, 38, 42, 43, 44], the inter-beat intervals in PPG signals (Figure 2, bottom) are correlated to users' emotions and affects. In addition, facial expressions captured from the front camera are also a rich source of users' emotions and affective states [15, 28].

We built RBF-kernel SVMs using the implicitly captured PPG signals and facial expressions to infer viewers' emotional responses to videos. Different from previous approaches in MOOC learning [35, 36, 42, 43, 44] on unmodified smartphones, there are at least two unique challenges when inferring emotional responses to mobile advertising. First, a video ad is much shorter (e.g. 15 - 60 seconds) than a tutorial video clip in MOOCs (e.g. 3 - 30 minutes), demanding higher sensitivity; Second, advertisers care about viewers' emotions such as "like" elicited by an advertisement and its potential to go "viral" (i.e. willingness to reshare [29]), while instructors in MOOCs pay more attention to learners' engagements, confusions

[42, 43], mind wandering [35], divided attention [44] and perceived difficulties [36] in learning. Moreover, AttentiveVideo uses both PPG signals and facial expressions for emotion detection, while AttentiveLearner only leverages PPG signals.

In this paper, we pose the following research questions:

- Can AttentiveVideo infer viewers' emotional responses in very short video clips?
- How do different modalities contribute to the detection of the emotions and affective states elicited by video advertising?
- Whether and how much can we reduce the required covering time while maintaining a high prediction accuracy?

In the following sections, we conduct a user study to answer these three questions.

USER STUDY

Experimental design

Previous work [15, 28, 31, 38] annotated experimental advertisements separately, i.e. users were required to watch an ad directly and annotate it before watching the next one. However, in real world scenarios, ads are usually grouped together and embedded into host video contents, e.g. movies and TV shows. In this research, we study video advertisements embedded in a real-world setting, i.e. an episode of a popular TV series (The Big Bang Theory). The episode has 3 embedded advertising slots and can be accessed freely on the official website². We only replace the original ads with our experimental ads while keeping the advertising positions unchanged.

We selected 12 video ads for the following brands: Ameriquet, Coca Cola, Doritos, Extra Gum, Guinness, Johnson & Johnson, One Main, Pepsi, Straight Talk, Township, Verizon, and Volkswagen. The mean length of the ads was 30.17s ($\sigma = 0.69$). The ads were chosen because they presented a range of affective states (i.e. humor, warmth, and neutral). It is worth noting that humor and warmth are important indicators of advertisement's effectiveness [1, 28].

Participants and Apparatus

We recruited 24 participants (13 females) from a local university for this study. The average age was 25.58 ($\sigma = 3.01$). Participants watched movies and TV shows on a regular basis (21 watched weekly and 3 watched monthly). Only one participant had not used mobile devices for video consumption.

Our experiment was completed on a Nexus 5 smartphone with a 4.95-inch, 1920 x 1080 pixel display, 2.26 GHz quad-core Krait 400 processor, running Android 5.0. It has an 8 megapixel back camera with a LED flash.

² http://www.cbs.com/shows/big_bang_theory

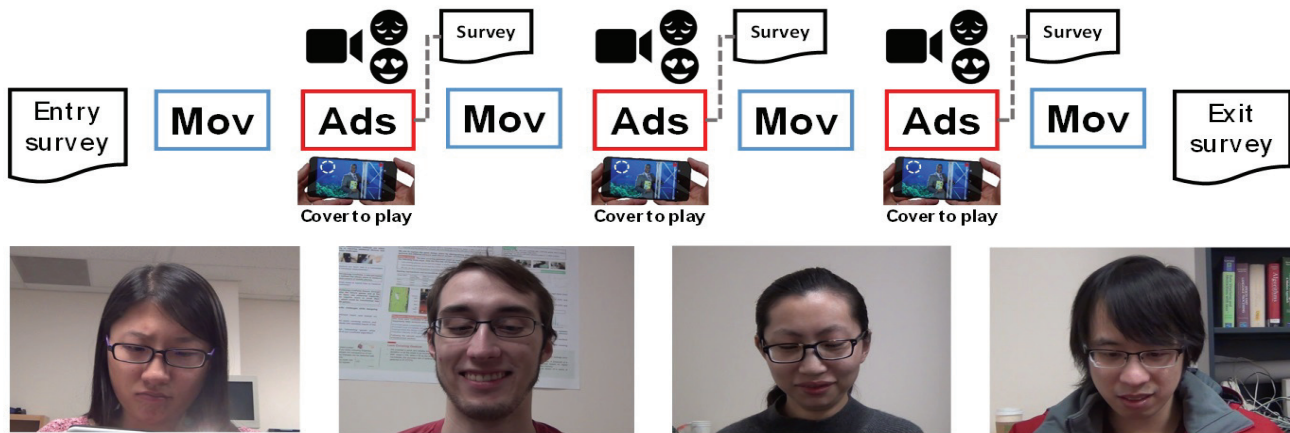


Figure 4. Experimental procedure (the top row) and sample facial images captured during the user study.

It is worth noticing that turning on two cameras in a smartphone in preview mode imposes major challenges in both hardware and software design. First, not all smartphones today allow the concurrent streaming of both the front camera and back camera at the same time, due to restrictions in camera firmware and the memory access architecture. Based on our experiments, only Google Nexus 6, Samsung Galaxy S4/S5, and Amazon Fire Phone are capable of turning on two cameras in preview mode at the same time. We hope by demonstrating the potential of collect PPG signals and facial expressions in parallel, more smartphone manufactures will start supporting such capabilities in the future; Second, it is critical to write efficient multi-thread functions to handle the video playback, PPG sensing from the back camera, and FEA from the front camera running on different physical processor cores to achieve real-time signal processing; Third, the write bandwidth for the external storage (i.e. the flash memory) of today's smartphones is insufficient for saving two video streams in real time. As a result, although we can achieve real-time signal processing for both cameras on a Nexus 6 at runtime, we need to use a second device to save the two video streams for offline training and benchmarking. To directly compare multiple algorithms offline, we use a separate camcorder as the front camera to save the original video stream for post-hoc analysis. We have successfully implemented the real-time parallel video processing algorithms of AttentiveVideo on a Nexus 6.

Procedures

Figure 4 illustrates the procedure of this study. A participant signed a consent form and answered a demographic survey. Then, the participant took a training session by watching a video consisting of two movie trailers and two embedded advertising slots with two ads in each slot. During this training session, the camcorder was adjusted to capture the participant's face. While watching non-advertised content (the movie trailers), the participant used normal on-screen gestures to play the video and the

camcorder was turned off. While watching the ads, the participant needed to cover the back camera lens to play and the camcorder was turned on to record the participant's facial expressions. After watching an advertising slot, containing two ads, the participant answered a subjective survey for each of the two ads before continuing to the non-advertised content. The participant took a short break before proceeding to the formal study session which has the same format as the training session. In the formal study, the participant watched an episode of The Big Bang Theory with three embedded advertising slots, each slot contained four ads. At the end of the study, the participant rated the usability of AttentiveVideo and ranked the six most liked ads. Each participant received a \$10 gift card after completing the study.

Category	Metric	Question
Attention	Attention	I paid sufficient <i>attention</i> to the entire ad
	Recall	I can <i>recall</i> major details in this ad
Engagement	Like	Please choose the 6 ads in this study that you <i>liked</i> best and rank them accordingly (1: most liked; 6: least liked)
	Rewatch	I'm interested in <i>watching</i> the ad <i>again</i> in the future
	Share	I found something special in the ad and want to <i>share</i> it with my friends
Sentiment	Touching	I found the ad <i>touching</i>
	Amusing	I found the ad <i>amusing</i>
	Valence	Self-Assessment Manikin
	Arousal	Self-Assessment Manikin

Table 1. Nine dimensions of emotional response measure.

Data Collection and Processing

Evaluation Metrics

Participants responded to six emotions related questions on the effectiveness of each advertisement, i.e. *Attention*, *Share*, *Touching* (feeling moved), *Rewatch*, *Recall*, and *Amusing*. We also used the Self-Assessment Manikin (SAM) [32] to collect responses for 2 continuous emotional metrics, i.e. *Valence* and *Arousal*. These ratings are in a 7-point Likert scale format (1: highly disagree; 7: highly agree). Figure 5 shows the average rating of each item for the 12 ads. On average, the emotional responses to experimental stimuli (ads) were diverse, which implied that the selected ads cover a wide range of advertisement's effectiveness dimensions. Some were highly rated for a single item while receiving low ratings on other items, e.g. ExtraGum received high ratings on the Touching metric but low ratings on others. In addition, participants rated the Like metric by ranking the 6 most liked ads at the end of the study. In total, we collected measures of 9 emotional states for each advertisement (Table 1).

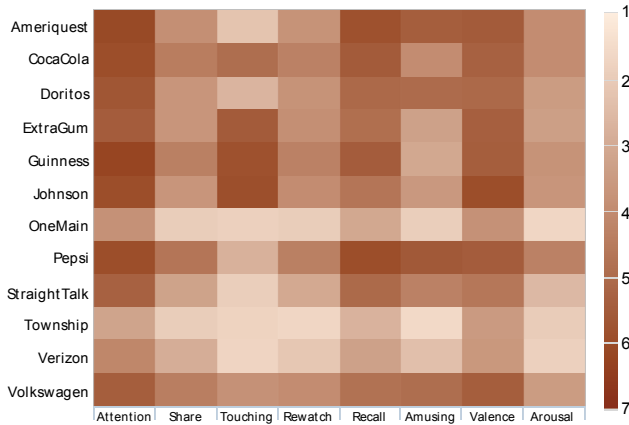


Figure 5. Average scores of each ad across 8 metrics on a 7-point Likert scale or Self-Assessment Manikin.

PPG Signal

AttentiveVideo extended the *LivePulse* algorithm [14] to identify the inter-beat intervals (NN) from PPG signals. LivePulse is a 6-step peaks/valleys counting heuristic using adaptive thresholding to remove outliers. The NN intervals are calculated as the distances between zero-crossing points which are interpolated from the identified peaks/valleys. The algorithm also uses a post-processing method taking into account NN intervals during the previous 5 seconds to remove noisy NN intervals. In this study, we further smooth the extracted NN intervals using the moving average window function and resample the intervals to 20Hz.

Similar to previous work [35, 36, 42], we extract 10 dimensions of HRV-related features from the NN intervals: 1) AVNN (average heart rate); 2) SDNN (temporal standard deviations of heart beats); 3) rMSSD; 4) pNN5 (percentage of adjacent heart beats with a difference longer than 5 ms); 5) pNN10; 6) pNN20; 7) SDANN; 8) SDNNIDX; 9)

SDNNIDX; 10) MAD (median absolute deviation). Since the duration of an ad is relatively brief when compared with a MOOC tutorial video, we replaced the common pNN50 feature with pNN5, pNN10, pNN20. For each participant, all features were rescaled to [0, 1].

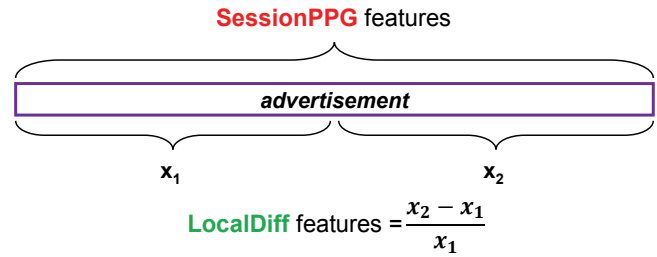


Figure 6. Extracting SessionPPG and LocalDiff features.

For each advertisement, the 10 dimensions of PPG features above were extracted in two different settings, i.e. the session feature set (SessionPPG) and the local difference feature set (LocalDiff). SessionPPG includes 10 dimensions of HRV features from an entire ad. LocalDiff extracted 10 dimensions of HRV features from both the first and the second half of the ad and then calculated their relative differences. Figure 6 illustrates how SessionPPG and LocalDiff were extracted from an ad. While SessionPPG had been successfully used in previous research [42], to the best of our knowledge, we are the first to define and use LocalDiff features on PPG signals.

Facial Expression

AttentiveVideo used *Affdex SDK* [2] to analyze facial expressions from the recorded clips. The means and standard deviations of 24 output features (15 action units and 9 emotions) were extracted for each advertisement in its entirety. In our preliminary analysis, using the top 10 features, chosen by Weka's InfoGain feature selection, led to a 2%-15% accuracy improvement in detecting emotional states when compared to using all 48 features. Therefore, we used the top 10 features for the experiment analysis. From now on, the facial expression features are referred as FEA features. Figure 2 shows four examples of faces detected by Affdex SDK with 26 feature points (landmarks).

Facial expression data from five participants were excluded from the follow-up analysis because the algorithm could not locate facial landmark points reliably from these participants.

Multimodal Features

To build multimodal systems, we used the feature fusion [10] approach to combine PPG signal features and FEA features. We only used the top 10 multimodal features selected by Weka's InfoGain to minimize the curse of dimensionality. I.e., the multimodal systems used the top 10 features of SessionPPG + FEA feature or the top 10 features of LocalDiff + FEA features.

Metric	FullDS				ExtremeDS			
	SessionPPG		LocalDiff		SessionPPG		LocalDiff	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Like	66.51%	0.31	66.03%	0.32	74.39%	0.47	72.63%	0.43
Attention	67.46%	0.33	64.59%	0.28	68.60%	0.35	73.33%	0.47
Share	66.99%*	0.32	62.20%	0.21	73.86%	0.45	73.51%	0.40
Touching	64.59%	0.26	66.99%	0.33	74.04%	0.45	70.70%	0.42
Rewatch	64.59%	0.27	63.64%	0.25	68.42%	0.36	70.18%	0.39
Recall	65.07%	0.29	64.11%	0.27	78.25%	0.54	73.16%	0.47
Amusing	64.59%	0.29	66.51%	0.32	68.42%	0.38	74.21%*	0.48
Valence	64.59%	0.28	66.03%	0.32	65.96%	0.32	70.70%	0.42
Arousal	62.20%	0.21	63.64%	0.25	66.84%	0.33	73.33%*	0.46

Table 2. Accuracy and Kappa of PPG signals (SessionPPG and LocalDiff) across 9 metrics. * indicates marginal differences ($p < 0.1$) between SessionPPG and LocalDiff models.

Model Building

From the annotated data, the emotional response detection task can be considered as a regression or ranking problem where our machine learning models predict a participant’s rating value. However, to evaluate the feasibility of AttentiveVideo in this pilot study, we start with binary classifiers which detect if a participant has a specific emotion for an ad. In other words, we build a binary classifier for each emotional state, e.g. Like or not Like and Amusing or not Amusing. A similar approach has been used in previous work [28, 29], where binary classifiers were evaluated instead of regression or ranking models.

To build the binary classifiers, we re-annotated the dataset using participants’ ratings. Following Greenwald et al. [13], for each metric, we sorted participants’ ratings and used the average rating of each ad in the dataset as a tie breaker. For example, let participant S1 watch 8 ads (a, b, c, d, e, f, g, h) and her “Like” ratings of these ads are (a, 1) (b, 2) (c, 6) (d, 7) (e, 3) (f, 6) (g, 7) (h, 3). Note that in this example, there are three ties (e = h, c = f, d = g). Given the average ratings of tied videos in the dataset satisfy $e > h$, $f > c$, and $g > d$ then S1’s ratings will be sorted as (a, 1) (b, 2) (h, 3) (e, 4) (c, 5) (f, 6) (d, 7) (g, 8). From this re-annotated dataset, we select the top 50% ads as positives (a, b, h, e) and the other 50% ads as negatives (c, f, d, g). We called this balanced dataset *FullDS* because it makes use of all data. We also created another dataset, called *ExtremeDS*, by selecting the top 25% ads as positives (a, b) and the bottom 25% ads as negatives (d, g). The *ExtremeDS* discarded weak (mid-ranked) emotional responses and only kept strong emotional responses while the *FullDS* kept both strong and weak responses. A similar approach has been done by McDuff et al. [29], where the authors only reported performance on a dataset having neutral and mild (weak) responses to ads removed. In this paper, we showed that comparing performance on both *FullDS* and *ExtremeDS* can reveal interesting insights of both modalities.

We built *user-independent* models and utilized the leave-one-subject-out cross-validation method for evaluation. RBF-kernel SVMs were chosen because they gave good performance with PPG signals [35, 36, 42, 43, 44] and FEA features [29]. The gamma and trade-off margin size hyperparameters were optimized using grid search in the range of [0.5, 1.7] with the step size equals to 0.2. The best hyperparameter set is considered as having the best cross-validation performance.

We also evaluated the prediction accuracies after discarding different portions of the signals at the beginning of each advertisement, called *starting offsets*, for two reasons. First, carry-over effects can occur when participants watched four ads continuously in an advertising slot. Discarding a portion of the signal from the beginning of each ad can prevent emotions from the previous ad propagating to the current ad. Second, we want to reduce user’s effort when covering the back camera lens while maintain the system’s accuracy. To achieve the best possible performance, we treated the duration of discarded signals as a hyper-parameter and tested the starting offsets from 2s to 18s, at the stride of 2s.

RESULTS AND ANALYSIS

Usability feedback

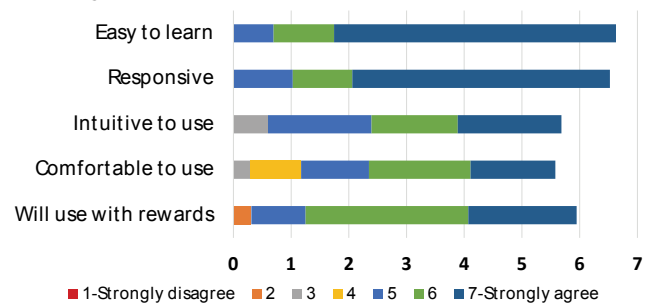


Figure 7. Subjective feedback of AttentiveVideo’s usability.

As shown in Figure 7, participants found AttentiveVideo easy to use ($\mu = 6.63$, $\sigma = 0.67$), responsive ($\mu = 6.53$, $\sigma =$

0.75), intuitive ($\mu = 5.68$, $\sigma = 1.22$), and comfortable ($\mu = 5.58$, $\sigma = 1.18$). Some positive comments were: ‘Good responsive, easy to use’, ‘Covering the lens of the back camera is natural when I hold the phone’, and ‘I don’t need to touch the screen to pull the menu icon to control’. However, participants were not fully satisfied with some aspects of AttentiveVideo. For example, integrating more advanced video controls (‘can adjust brightness and audio volume at the same time’) or having concerns about the flash light usage (‘Turning on the flash light all the time may use the battery faster’, ‘if the video is long, it may be hot to fingers’). We believe both richer video control mechanisms and better power management features can be added in future implementations.

Signal Quality

We analyzed the signal quality obtained by AttentiveVideo using a method that is similar to Xiao and Wang [42]. We used a 5-second NN interval signal window. The signal window is classified as good if at least 80% of the NN intervals are within $\pm 25\%$ of the window’s median. In 82.46% of 57 advertising slots (19 participants \times 3 slots), more than 89% of the signals were high quality. This suggests that AttentiveVideo can collect high quality signals from unmodified smartphones. Figure 8 illustrates the PPG signal quality captured by AttentiveVideo in the first advertising slot of six participants.

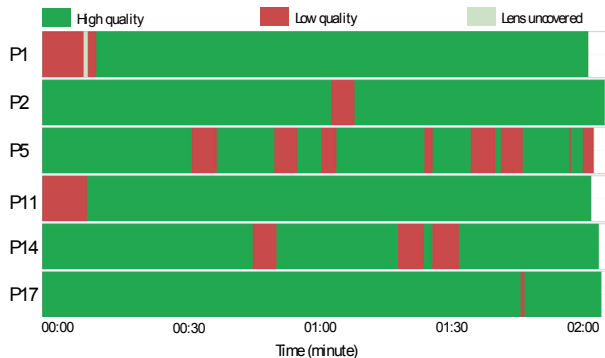


Figure 8. PPG signal quality of six participants while watching the first advertising slot.

Figure 9 shows HRV spectrograms (normalized amplitude) from the affect responses for the least touching (top row) and most touching (second row) ads of five participants. The HRV spectrograms were computed by calculating the power spectral density from NN intervals. For each ad, because of its short duration, we used a 20-second sliding window with half-second increments. Previously, McDuff [27] found that the high frequency (HF) power decreased under a stress condition. Figure 9 shows lower HF power in the HRV spectra for those watching the least touching ads

compared to watching the most touching ads. This indicates those participants felt less stressed (or more relaxed) when watching the most touching ad than watching the least touching ad.

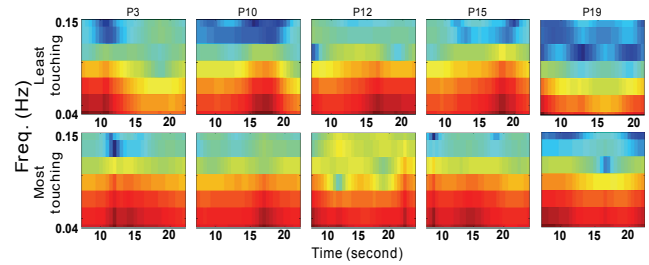


Figure 9. HRV spectrograms of the least touching ads (top row) and the most touching ads (second row) from five participants: P3, P10, P12, P15, and P19.

Emotional Response Detection

Table 2 shows the performance of two PPG feature types (SessionPPG and LocalDiff) across nine emotional response metrics. All experimental models outperformed the random classifier (Accuracy = 50.00%, Kappa = 0.00). Comparing the two PPG feature types, we found no significant differences between SessionPPG and LocalDiff. However, there were marginal differences in FullDS (Share: $t(18) = 1.6$, $p < 0.1$) and in ExtremeDS (Amusing: $t(18) = -1.37$, $p < 0.1$ and Arousal: $t(18) = -1.43$, $p < 0.1$). The differences implied that the two PPG feature types would complement each other in some emotional metrics. To evaluate PPG-based models with FEA-based models and multimodal models, from now on, we only report the best PPG feature type (either SessionPPG or LocalDiff) in each emotional response metric.

Figure 10 shows the accuracy of the PPG-based, FEA-based, and multimodal models in ExtremeDS and FullDS. On average, in ExtremeDS, our PPG-based models and FEA-based models achieved comparable performance with previous work using the same modality. In particular, the average performance of our PPG-based models is Accuracy = 73.59% (Kappa = 0.46). This accuracy is comparable with the performance of AttentiveLearner when it predicts boredom (78.56%) [42], confusion (81.96%) [42], and mind wandering events (71.22%) [35] in mobile learning. On the other hand, the average Accuracy of our FEA-based models is 70.51% (Kappa = 0.40). This performance is comparable with the accuracy in McDuff et al. [28], where the RBF SVMs’ accuracies for Like and Rewatch metrics are below 70%. Combining PPG features and FEA features, the multimodal models achieved the average performance as Accuracy = 61.46%, Kappa = 0.21 in FullDS and Accuracy = 73.43%, Kappa = 0.46 in ExtremeDS.

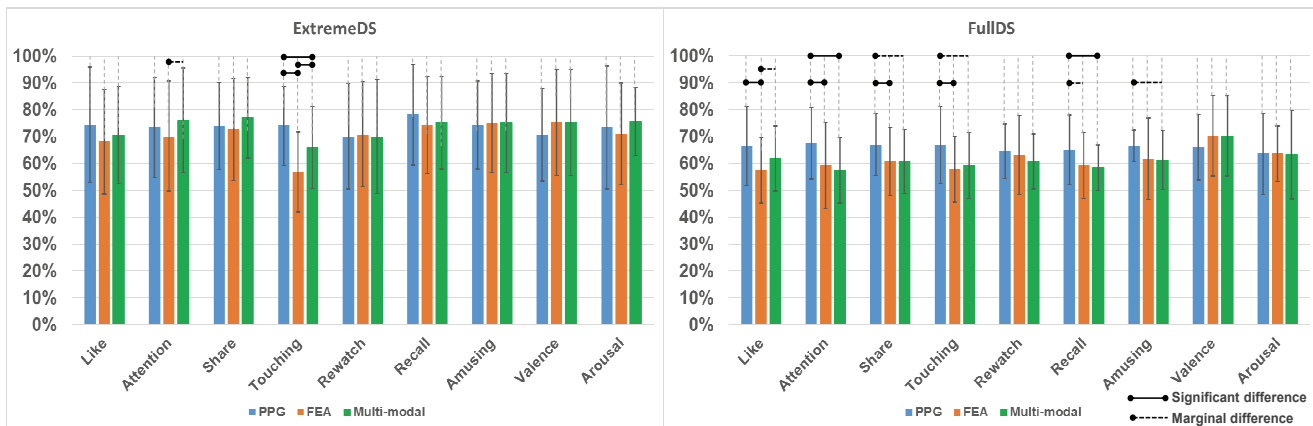


Figure 10. Accuracies of PPG-based, FEA-based, and multimodal models across 9 emotional metrics in ExtremeDS and FullIDS.

We found that discarding weak (mid-ranked) responses improve the models' performance. The accuracies of PPG-based models were significantly different in all emotional metrics between FullIDS and ExtremeDS, except the Valence metric ($t(18) = -0.29$, $p = 0.385$). Similarly, the accuracies of FEA-based models were significantly different in all metrics between FullIDS and ExtremeDS, except the Touching metric ($t(18) = 0.26$, $p = 0.4$). These insignificant differences suggested that PPG-based models would not be appropriate for the Valence metric and FEA-based models would not be a good choice for the Touching metric. Further investigations confirmed this assumption. In Figure 10, PPG-based models (FullIDS: 66.98%, ExtremeDS: 74.03%) outperformed FEA-based models (FullIDS: 57.89%, ExtremeDS: 56.84%) in the Touching metric while FEA-based models (FullIDS: 70.33%, ExtremeDS: 75.26%) achieved higher accuracy than PPG-based models (FullIDS: 66.03%, ExtremeDS: 70.70%) in the Valence metric. These results suggest that these modalities are complementary. By using the appropriate modality for each emotional metric, our framework can gain better overall performance.

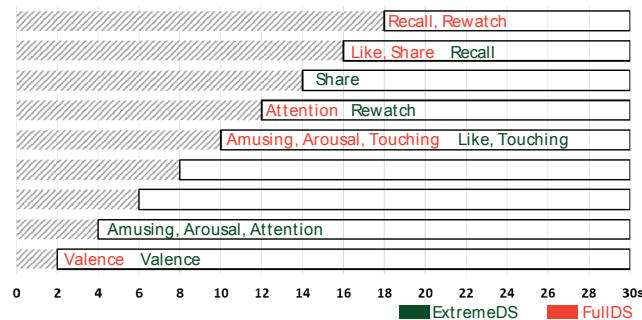


Figure 11. Starting offsets of nine metrics in FullIDS (orange) and ExtremeDS (green).

Starting Offset

Figure 11 shows that the starting offsets are different across measures of emotional states and datasets. Most measures allow a starting offset of up to 10s, the exceptions being

Amusing (4s), Arousal (4s), Attention (4s), and Valence (2s) in ExtremeDS. In FullIDS, Valence also has a starting offset of 2s. This finding implies that AttentiveVideo can further reduce user effort by allowing ad watching freely for the first 10s and then covering the lens to play. Our results showed that the Touching measure allows a larger starting offset (10s in both FullIDS and ExtremeDS) than Amusing (4s in ExtremeDS and 10s in FullIDS). This difference could be explained in terms of *key events*' positions in each ad. It only takes a funny scene to make viewers laugh. Therefore, the key event in amusing ads could be at any position. On the other hand, touching ads require some time to setup a context where viewers can feel moved by the story. The key events of touching ads would appear at the end rather than the beginning of an ad. As a result, the Touching detection models will not be hurt when skipping a portion of the signal from the beginning of an advertisement.

DISCUSSIONS

We found that PPG-based models work better when predicting *subtle* emotional responses, while FEA-based models work better with *strong* emotional responses. In FullIDS (Figure 10, right), PPG-based models outperformed FEA-based models significantly or marginally on 5 measures: Like ($t(18) = 1.97$, $p < 0.05$), Attention ($t(18) = 1.90$, $p = 0.05$), Share ($t(18) = 1.95$, $p = 0.05$), Touching ($t(18) = 2.92$, $p < 0.005$), and Recall ($t(18) = 1.48$, $p < 0.1$). However, in ExtremeDS (Figure 10, left), PPG-based models performed significantly better than FEA-based for only one emotional measure, i.e. Touching ($t(18) = 2.94$, $p < 0.005$). In other words, FEA-based models gained better performance, relative to PPG-based models, after discarding weak emotional responses (ExtremeDS). On the other hand, PPG-based models still maintain good performance with weak emotional responses in FullIDS. McDuff [27] has a similar observation: some participants did not express much facial behavior in the stress condition, but there were changes in their physiological responses, i.e. blood volume pulse, respiration, and electrodermal activity, compared to the baseline condition.

PPG-based models performed consistently better than FEA-based models for the Touching measure (FullDS: $t(18) = 2.92$, $p < 0.005$ and ExtremeDS: $t(18) = 2.94$, $p < 0.005$). This result suggests that our participants might not feel moved by the experimental ads enough to make explicit facial gestures. As a result, PPG-based models can capture the dynamic changes from PPG signals which may not be easily captured from facial expressions.

Our current research on AttentiveVideo has only scratched the surface of making user interfaces emotion-aware. There are still important directions worth exploring in the future.

First, we focused on video advertisements around 30 seconds long. We believe AttentiveVideo should have reasonable performance for longer ads considering the successes in detecting cognitive and affective states from longer tutorial videos [35, 36, 42, 43, 44]. However, it is still an open problem for even shorter ads, e.g. 5 - 15 seconds, considering that the temporal resolution of our PPG-based models is bounded by the window size for extracting HRV features. In fact, time-domain HRV features aim to track the dynamics in NN intervals within a signal window. Analyzing a single NN interval value would not give any interesting finding. In comparison, FEA-based algorithms can make instant predictions based on a single video frame of the viewer's face. We hypothesize that FEA-based models will have a higher weight than PPG-based models for shorter video clips.

Second, we used *feature fusion* [10] to create our models combining PPG and FEA signals. These models still have room for improvements. For example, we applied the same starting offset for these two modalities in our experiments even though each modality could have different optimal starting offsets. It would be interesting to explore *model fusion* in the future, considering that a model fusion approach may take the best from each unimodal system rather than introducing more constraints in feature normalization or hyperparameter tuning. We will explore voting mechanisms, e.g. average vote and weighted majority vote [17], for the model fusion approach. Moreover, it would be interesting to explore dimensionality reduction techniques in addition to feature selection techniques in follow-up experiments.

CONCLUSIONS

We presented AttentiveVideo, an intelligent mobile player that collects two rich sets of affect data using different modalities and infers viewers' emotional responses towards video advertisements on unmodified smartphones. AttentiveVideo can predict viewers' *attention, engagement, and sentiment* towards advertisements *via a combination of implicit PPG sensing and FEA on today's smartphones*. In a 24-participant study, we found that AttentiveVideo achieved good accuracy on a wide range of emotional states (best average accuracy = 73.59%, kappa = 0.46 across 9 measures). We also found that the PPG sensing channel and the FEA technique are complimentary. While FEA works

better for strong emotions (e.g., joy and anger) and is able to give instant predictions, the PPG channel is more informative for subtle responses or emotions but requires more time (several seconds) to make predictions. Furthermore, a multimodal system combining both PPG sensing and FEA achieved better accuracy for detecting Attention, Like, and Touching compared to FEA models.

AttentiveVideo can help advertisers to have a richer and fine-grained understanding of users' emotional responses towards video advertisements. AttentiveVideo can also help viewers to enjoy more high quality video materials for free via subsidized video ads.

ACKNOWLEDGEMENTS

We thank Xiang Xiao, Xiangmin Fan, Carrie Demmans Epp, Jesse Davis, and the anonymous reviewers for the help and constructive feedback. This research is in-part supported by an RDF from the Learning Research and Development Center (LRDC) at the University of Pittsburgh, and a generous gift from Byte Dance Telecommunications Co. Ltd.

REFERENCES

1. David A. Aaker, Douglas M. Stayman, and Michael R. Hagerty. Warmth in advertising: Measurement, impact, and sequence effects. *Journal of Consumer Research* 12, no. 4 (1986): 365-381.
2. Affectiva. Infusing Affectix emotion sensing and analytics anywhere. 2009. Retrieved October 14, 2016 from <http://www.affectiva.com/solutions/apis-sdks/>
3. Jeremy N. Bailenson, Emmanuel D. Pontikakis, Iris B. Mauss, James J. Gross, Maria E. Jabon, Cendri AC Hutcherson, Clifford Nass, and Oliver John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies* 66, no. 5 (2008): 303-317.
4. Hila Becker, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Bo Pang. Context transfer in search advertising. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 656-657. ACM, 2009.
5. Paul D. Berger, and Nada I. Nasr. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing* 12.1 (1998): 17-30.
6. Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems*, pp. 55-60. Springer International Publishing, 2014.
7. Andrei Z. Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. Search advertising using web relevance

- feedback. In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1013-1022. ACM, 2008.
8. Rafael A. Calvo, and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1, no. 1 (2010): 18-37.
9. Xiu Y. Chen, and Zary Segall. XV-Pod: An emotion aware, affective mobile video player. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol. 3, pp. 277-281. IEEE, 2009.
10. Sidney K. D'Mello, and Arthur Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, no. 2 (2010): 147-187.
11. Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American economic review* 97, no. 1 (2007): 242-259.
12. Roman Ganhör. ProPane: fast and precise video browsing on mobile phones. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, pp. 20. ACM, 2012.
13. Mark K. Greenwald, Edwin W. Cook, and Peter J. Lang. Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology* 3, no. 1 (1989): 51-64.
14. Teng Han, Xiang Xiao, Lanfei Shi, John Canny, and Jingtao Wang. Balancing accuracy and fun: Designing camera based mobile games for implicit heart rate monitoring. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 847-856. ACM, 2015.
15. Richard L. Hazlett, and Sasha Y. Hazlett. Emotional response to television commercials: Facial EMG vs. self-report. *Journal of Advertising Research* 39 (1999): 7-24.
16. Nigel Hollis. Ten years of learning on how online advertising builds brands. *Journal of advertising research* 45, no. 2 (2005): 255-268.
17. M. S. Hussain, Hamed Monkaresi, and Rafael A. Calvo. Combining classifiers in multimodal affect detection. In Proceedings of the Tenth Australasian Data Mining Conference-Volume 134, pp. 103-108. Australian Computer Society, Inc., 2012.
18. Andrew H. Kemp, and Daniel S. Quintana. The relationship between mental and physical health: insights from the study of heart rate variability. *International Journal of Psychophysiology* 89, no. 3 (2013): 288-296.
19. Annie Lang. Involuntary attention and physiological arousal evoked by structural features and emotional content in TV commercials. *Communication Research* 17, no. 3 (1990): 275-299.
20. Kuang-chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. Estimating conversion rate in display advertising from past performance data. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 768-776. ACM, 2012.
21. Lusong Li, Tao Mei, and Xian-Sheng Hua. GameSense: game-like in-image advertising. *Multimedia Tools and Applications* 49, no. 1 (2010): 145-166.
22. Yuheng Li, Yiping Zhang, and Ruixi Yuan. Measurement and analysis of a large scale commercial mobile Internet TV system. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pp. 209-224. ACM, 2011.
23. Zhenyu Li, Gaogang Xie, Mohamed Ali Kaafar, and Kave Salamatian. User behavior characterization of a large-scale mobile live streaming system. In Proceedings of the 24th International Conference on World Wide Web, pp. 307-313. ACM, 2015.
24. Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, no. 1 (2003): 76-80.
25. Ritu Lohtia, Naveen Donthu, and Edmund K. Hersherberger. The impact of content and design elements on banner advertising click-through rates. *Journal of advertising Research* 43, no. 04 (2003): 410-418.
26. John M. Murphy, *Branding: A key marketing tool*. Springer, 1987.
27. Daniel McDuff. 2014. *Crowdsourcing affective responses for predicting media effectiveness* Ph.D Dissertation. Massachusetts Institute of Technology, Cambridge, MA.
28. Daniel McDuff, Rana El Kaliouby, Jeffrey F. Cohn, and Rosalind W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing* 6, no. 3 (2015): 223-235.
29. Daniel McDuff, Rana El Kaliouby, Thibaud Senechal, David Demirdjian, and Rosalind Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing* 32, no. 10 (2014): 630-640.
30. Tao Mei, Xian-Sheng Hua, and Shipeng Li. VideoSense: A contextual in-video advertising system. *IEEE Transactions on Circuits and Systems for Video Technology* 19, no. 12 (2009): 1866-1879.

31. Anca Cristina Micu and Joseph T. Plummer. Measurable emotions: How television ads really work. *Journal of Advertising Research* 50, no. 2 (2010): 137-153.
32. Jon D Morris. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research* 35, no. 6 (1995): 63-68.
33. Suman Nath, Felix Xiaozhu Lin, Lenin Ravindranath, and Jitendra Padhye. SmartAds: bringing contextual ads to mobile apps. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pp. 111-124. ACM, 2013.
34. Patrick De Pelsmacker and Maggie Geuens. The advertising effectiveness of different levels of intensity of humour and warmth and the moderating role of top of mind awareness and degree of product use. *Journal of Marketing Communications* 5, no. 3 (1999): 113-129.
35. Phuong Pham and Jingtao Wang. AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In *International Conference on Artificial Intelligence in Education*, pp. 367-376. Springer International Publishing, 2015.
36. Phuong Pham and Jingtao Wang. Adaptive Review for Mobile MOOC Learning via Implicit Physiological Signal Sensing. In *Proceedings of the 2016 ACM on International Conference on Multimodal Interaction*, pp. 37-44. ACM, 2016.
37. Rosalind W Picard. *Affective computing*. Vol. 252. Cambridge: MIT press, 1997.
38. Patricia A. Stout, and John D. Leckenby. Measuring emotional response to advertising. *Journal of Advertising* 15, no. 4 (1986): 35-42.
39. The Interactive Advertising Bureau (IAB). US Internet Ad Revenue in 2015. Retrieved October 14, 2016 from <http://www.iab.com/news/us-internet-ad-revenues-hit-landmark-59-6-billion-in-2015/>
40. Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, no. 2 (2013): 153-163.
41. Yue Wu, Tao Mei, Nenghai Yu, and Shipeng Li. Accelerometer-based single-handed video browsing on mobile devices: design and user studies. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, pp. 157-160. ACM, 2012.
42. Xiang Xiao and Jingtao Wang. Towards Attentive, Bi-directional MOOC Learning on Mobile Devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 163-170. ACM, 2015.
43. Xiang Xiao and Jingtao Wang. Context and cognitive state triggered interventions for mobile MOOC learning. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 378-385. ACM, 2016.
44. Xiang Xiao, and Jingtao Wang. Understanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 35rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2017.
45. Xiang Xiao, Teng Han, and Jingtao Wang. LensGesture: augmenting mobile interactions with back-of-device finger gestures. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 287-294. ACM, 2013.
46. Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World wide web*, pp. 261-270. ACM, 2009.
47. YouTube. Analytics and Reporting APIs. 2016. Retrieved October 14, 2016 from <https://developers.google.com/youtube/analytics/v1/diagram/mets>
48. Jin-Kai Zhang, Cui-Xia Ma, Yong-Jin Liu, Qiu-Fang Fu, and Xiao-Lan Fu. Collaborative interaction for videos on mobile devices based on sketch gestures. *Journal of Computer Science and Technology* 28, no. 5 (2013): 810-817.
49. Weinan Zhang, Shuai Yuan, and Jun Wang. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1077-1086. ACM, 2014.